

THE LIMSI TOPIC TRACKING SYSTEM FOR TDT2001

Yuen-Yee Lo and Jean-Luc Gauvain

Spoken Language Processing Group

LIMSI-CNRS, BP 133

91403 Orsay cedex, FRANCE

{yylo,gauvain}@imsi.fr <http://www.limsi.fr/tlp>

OVERVIEW

- Development Conditions
- Unigram Tracker
- Stopping and Stemming
- Document Expansion
- Unsupervised Online Adaptation
- Experimental Results
- Conclusions

DEVELOPMENT CONDITIONS

- TDT2 corpus, Jan-Jun 1998
 - 90 000 documents, 42M words
 - 96 released topics
 - GE model training, Document expansion
- TDT3 development corpus, Oct-Dec 1998
 - 60 released topics
- Data sources: newswire, Broadcast News (BN)

UNIGRAM TRACKER

- Similarity measure: normalized log likelihood ratio between the topic model T and a general English model (GE)

$$S(story, T) = \frac{1}{L_d} \log \frac{\Pr(story|T)}{\Pr(story|GE)}$$

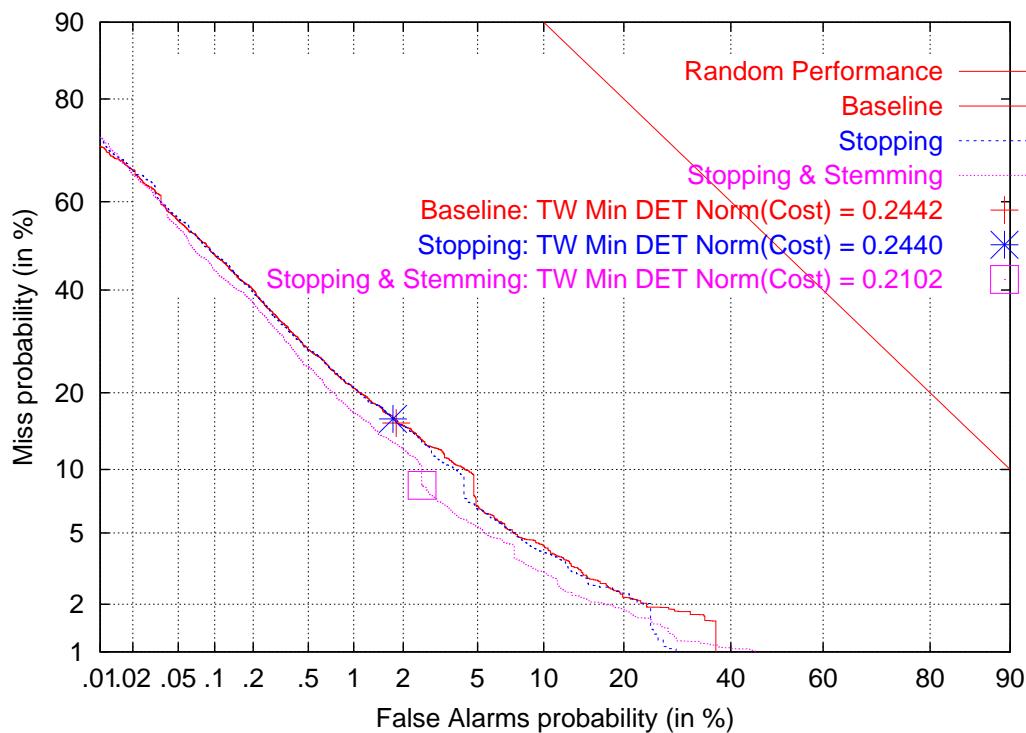
- $\Pr(\cdot|GE)$ is trained on the TDT2 corpus
- $\Pr(\cdot|T)$ is obtained by interpolating its ML unigram estimates with the general English model
- Similarity score is normalized by the story length (L_d).
- Similarity score $> t_d$ (the decision threshold), the story is identified on topic

STOPPING & STEMMING

- Stopping: 800 stop words
- Stemming: Porter stemmer with manual correction
- Stemmed lexicon with 38000 entries

Condition	Nt=1	Nt=4
	nwt+bnman manual bound.	nwt+bnasr auto. bound.
Baseline	0.2442	0.1728
Stopping	0.2440	0.1678
Stopping & Stemming	0.2102	0.1368

IMPACT OF STOPPING & STEMMING (Nt=1)

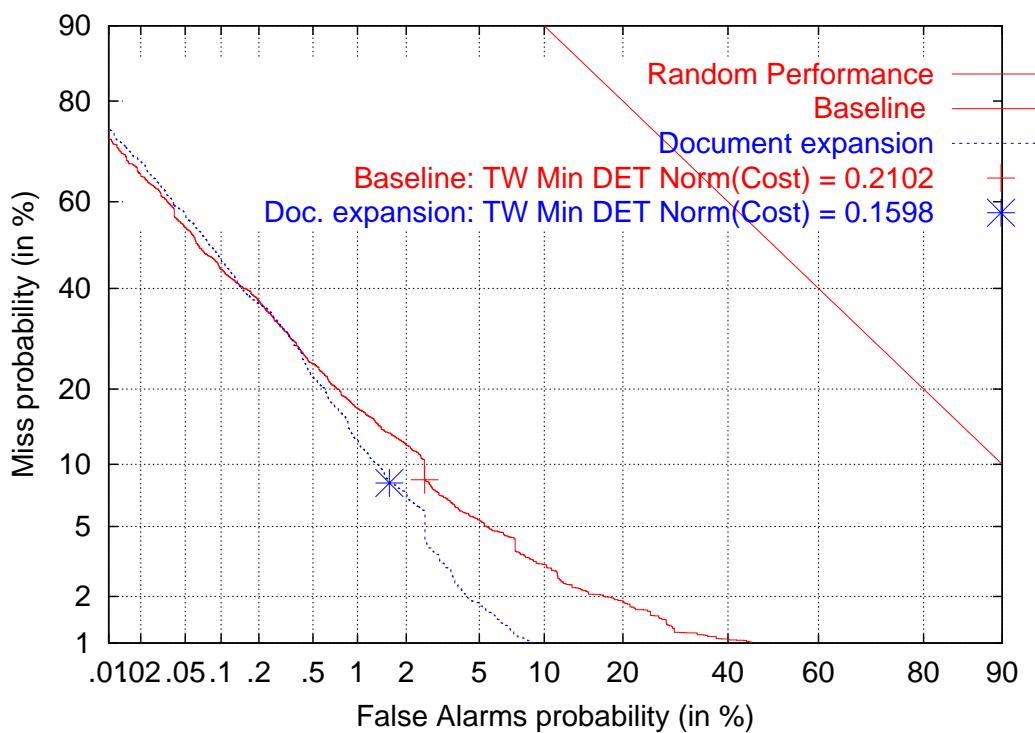


DOCUMENT EXPANSION

- Add related terms to the on-topic training stories
- Use query expansion technique developed for TREC SDR task
- OKAPI-based information retrieval system
- Text corpus: 42M words of TDT2 texts (New York Times, Los Angeles Times, Washington Post, Jan-Jun 1998)
- Add 25 terms with term frequencies proportional to their offer-weight
- Example, topic 30001 : Cambodian Government Coalition

Term	hun	sen	ranariddh	cambodia	prince	norodom	elect	khmer
Weight	2.27	2.06	1.96	1.83	1.26	1.15	1.14	1.12

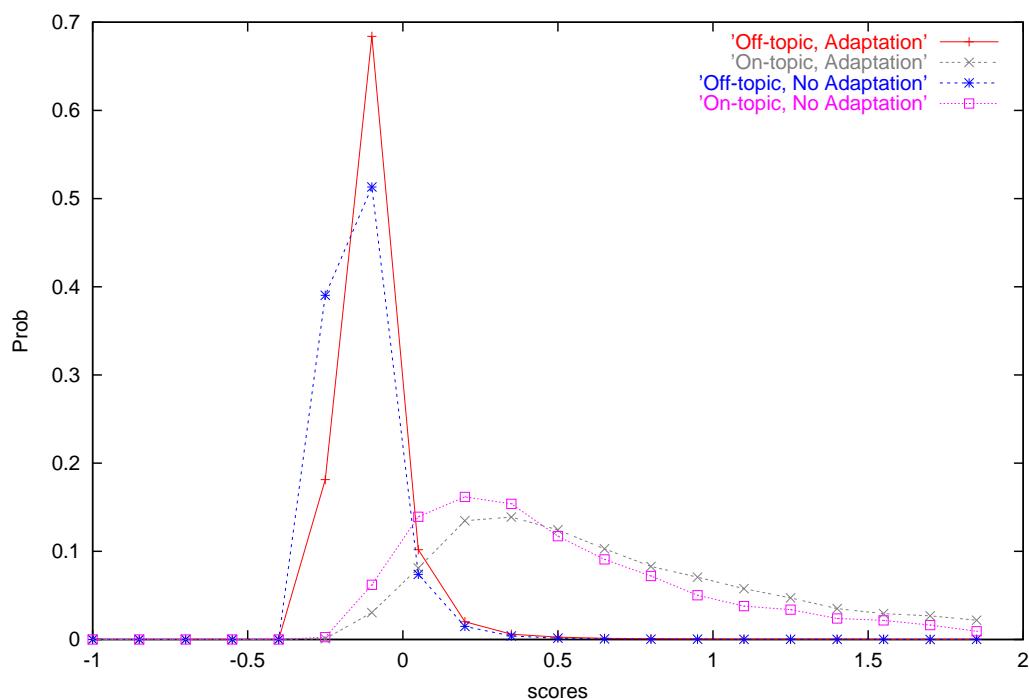
IMPACT OF DOCUMENT EXPANSION



UNSUPERVISED ONLINE ADAPTATION

- Topic model adaptation by adding incoming stories to training data
- Use only stories identified as on-topic by the system
- Score higher than an adaptation threshold th_a ($th_a \geq th_d$)
- Adaptation data weight:
 - Fixed (independent of the score)
 - Based on the similarity score
- Number of adaptation steps: unlimited

ON-OFF-TOPIC SCORE DISTRIBUTION



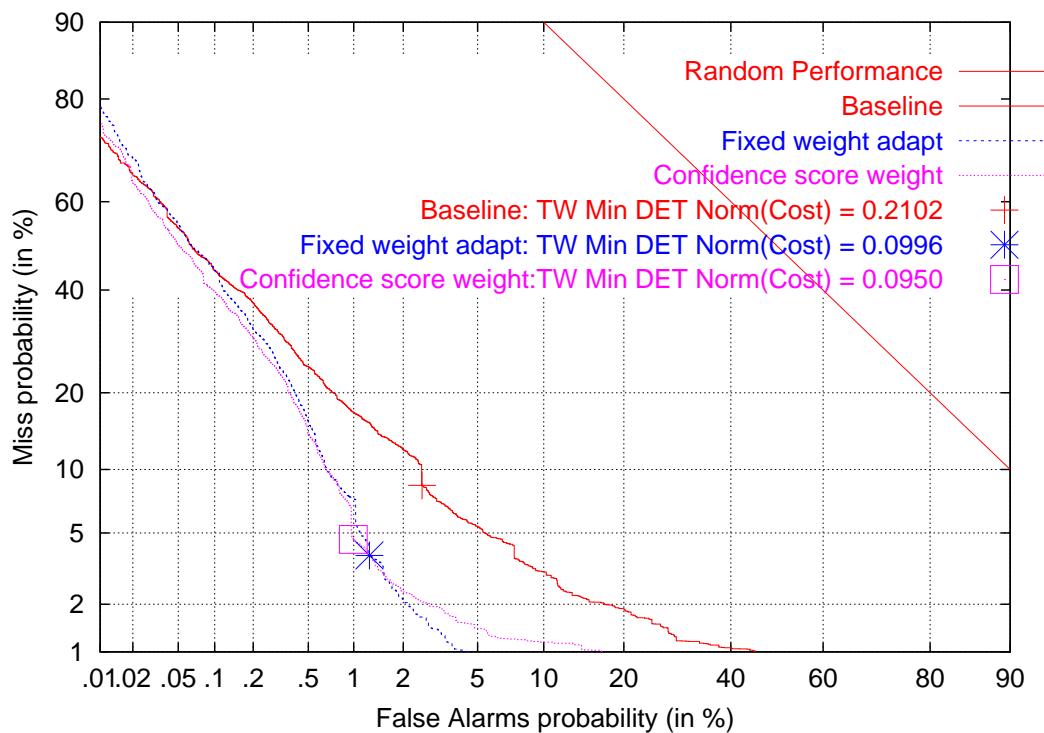
VARIABLE ADAPTATION WEIGHT

- Map similarity score to a confidence score using a piece-wise linear transformation

$$s(T, \text{story}) \rightarrow \Pr(T|\text{story})$$

- Train mapping on the TDT3 development data for each condition
- Use the confidence score as an adaptation weight
- Use all hypothesized on-topic stories for adaptation

DET WITH AND WITHOUT ADAPTATION



SUMMARY OF DEVELOPMENT RESULTS - NT1

Sources	nwt+bnman	nwt+bnasr	nwt+bnasr
Boundary	manual*	manual	auto
Baseline tracker	0.2102	0.2317	0.2271
Doc. expansion	0.1598	0.1780	0.1753
Fixed weight adapt.	0.0996	0.1089	0.1353
Confidence score adapt.	0.0950	0.1086	0.1337
Doc. exp. & conf. adapt	0.0947	0.1046	0.1281

* Primary condition

SUMMARY OF DEVELOPMENT RESULTS - NT4

Sources	nwt+bnasr	nwt+bnasr
Boundary	manual	auto
Baseline tracker	0.1288	0.1368
Doc. expansion	0.1256	0.1326
Fixed weight adapt.	0.1095	0.1143
Confidence score adapt.	0.0916	0.1111
Doc. exp. & conf. adapt	0.0946	0.1136

TDT2001 RESULTS

- 5 conditions, 7 submissions

- 2 system versions (LIMSI-1, LIMSI-2)

LIMSI-1: document expansion and unsupervised adaptation

LIMSI-2: unsupervised adaptation only

Nt	Sources	Boundaries	LIMSI-1	LIMSI-2
1	nwt+bnasr	auto	0.1797	-
1	nwt+bnasr	manual	0.1294	-
1	nwt+bnman	manual	0.1213	-
4	nwt+bnasr	manual	0.1415	0.1490
4	nwt+bnasr	auto	0.1842	0.1921

CONCLUSIONS

- First LIMSI participation in TDT
- Unigram tracker
- Document expansion and online adaptation
- Cost reduction for primary condition
 - Document expansion: 23%
 - Online adaptation: 54%
 - Expansion & adaptation: 55%
- TDT2001 for primary condition tracking cost: 0.1213